Michael Geiger
EM 212: Applied Data Science
Project #1 Revision
2/28/18

Executive Summary

For my data science project this semester, I would like to investigate how accurately you can predict whether or not the outcome of an NFL game will fall within the projected outcome (right winner predicted), the points that the winning team will win by (spread), and the combined points (over/under) for a specific game. These predictions will be based on historical betting data, NFL game matchups and outcomes, and various environmental conditions such as stadium location and weather conditions. I will also be merging in external population data to see how additional information affects the accuracy of this model. With the rise of sports gambling in all professional sports in the United States, a predictive model that can accurate prove/disprove projected betting odds for any game is an incredibly valuable tool. With the NFL being one of the most popular sports in the US, a predictive tool for personal-gambling purposes would be incredibly valuable for use in states where sports gambling is currently legal.

The main datasets I will be using were located on Kaggle's dataset for "NFL scores and betting data." Currently, there has been little activity revolving around this specific dataset on Kaggle. There have been 4 studies related to these data on Kaggle:

1. https://www.kaggle.com/twalters20/nfl-betting-model
2. https://www.kaggle.com/northofmanhattan/james-hoffman-final-project-part-2
3. https://www.kaggle.com/tobycrabtree/nflbetr
4. https://www.kaggle.com/northofmanhattan/james-hoffman-final-project-part-3

Studies 1 and 3 have both been used to build models to predict the percentage that the home team or away team will win in any given matchup. My proposed study is different because I will be focusing on whether or not the projected betting outcomes will come true, not just the percent chance that a team will win. Next, studies 2 and 4 are actually the same study, which focuses on trying to predict the outcome of the 2017 Super Bowl by aggregating season-long outcomes and stats from every game. My study ignores the Super Bowl and playoffs entirely and is not accumulating data for any-given season to make predictions on future outcomes of games. My model will also be different than all of these models because I will be strictly focusing on games played during or after the 2010 season, and each of these models uses a much larger historical database for their models. None also brings in external data to build their models, further ensuring the uniqueness of my work.

Kaggle provides three different datasets for this topic: NFL Stadium information, NFL Team information, and NFL game information. For my project, I will be conducting the majority of my analysis around the NFL game information dataset that contains the scores of all games from the '96-'97 season to the '17-'18 season. In this dataset, each row is associated with a specific game number, and the corresponding columns for each row are as follows: game date, season

year, season week, playoff/regular season game, home team name, away team name, home team score, away team score, team favorited to win, the predicted score differential, the predicted over-under line, the stadium, temperature, wind speed, humidity, and whether or not the stadium has a dome. The other two datasets contain bits of information that I plan on using to supplement this main dataset. For the NFL Stadium information, the dataset lists all the past stadiums at which games were played. I will be using this dataset to add columns to the main dataset for stadium location, stadium type, stadium surface, and elevation for each game. For the NFL Team information dataset, each NFL team to have played a game is listed along with each teams' ID, Conference, and Division. To include this supplemental information in the main dataset, I will need to reduce all Team Names and Stadium Names to reference numbers across all datasets to simplify the dataset-reconstruction process.

The focus of my model will be on games played in the modern era of the NFL, so I will be deleting games played before a certain date (likely before either 2000 or 2010). In addition, there is a decent amount of missing information in important columns, so I will need to supplement these gaps using public information. As described above, I will also likely need to reduce name-data to reference values. I will also need to account for team name changes over my selected time frame of data among other occurrences of redundant/misspelled information. I will also be reducing TRUE/FALSE values to a binary 1/0 system. All of this will make the overall coding process more simplified in the long run.

As far as building a predictive model goes, I will be likely training my classifier on the games played between 2010 and 2017, leaving the 2018 season to test my classifier's prediction accuracy. The outcome of my classifier will be given in terms of home and away team scores, indicating the winner of the game, information which I will then leverage against the predicted betting data for that game. In practice, users of this predictive model would be able to apply the model for predicting the outcomes of games being played on a week-to-week basis for the 2019-2020 NFL Season.

The link to the Kaggle page that contains the relevant datasets is located below:
https://www.kaggle.com/tobycrabtree/nfl-scores-and-betting-data/home

The link to the external data I will be merging is located here:
https://catalog.data.gov/dataset/500-cities-city-level-data-gis-friendly-format-845f9